

DOCUMENT RESUME

ED 121 810

TM 005 238

AUTHOR Gallagher, P. D.; Gay, L. R.
TITLE The Comparative Effectiveness of Tests Versus Written Exercises in a Competency-Based Research Course.
PUB DATE [Apr 76]
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)
EDRS PRICE MF-\$0.83 HC-\$1.67 Plus Postage
DESCRIPTORS *Comparative Analysis; Graduate Students; Higher Education; Performance Based Education; *Retention; *Student Testing; *Writing Exercises

ABSTRACT

The purpose of this study was to determine whether tests would result in greater retention of research concepts and procedures than written exercises. Students in an introductory educational research course were randomly assigned to one of three treatment conditions--(1) written exercises only, (2) short-answer tests, or (3) individual choice. A one-way analysis of variance (ANOVA) indicated that there was a significant difference among the three groups on a "surprise" final examination. Application of Scheffe's multiple comparison procedure indicated that students in the test condition retained significantly more than students in both the written exercise condition and choice condition. (Author/RC)

* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *

ED121810

THE COMPARATIVE EFFECTIVENESS OF
TESTS VERSUS WRITTEN EXERCISES IN
A COMPETENCY-BASED RESEARCH COURSE

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

P. D. Gallagher and L. R. Gay
Florida International University
Miami, Florida

Paper Presented at the American Educational
Research Association Convention
April, 1976

TM005 238

ABSTRACT

The purpose of this study was to determine whether tests would result in greater retention of research concepts and procedures than written exercises. Students in an introductory educational research course ($N = 126$) were randomly assigned to 1 of 3 treatment conditions - 1) written exercises only, 2) short-answer tests, or 3) individual choice. A one-way ANOVA indicated that there was a significant difference among the 3 groups on a "surprise" final examination. Application of Scheffé's multiple comparison procedure indicated that students in the test condition retained significantly more than students in both the written exercise condition and choice condition.

CONTENTS

	Page
INTRODUCTION	
Statement and Background of the Problem.....	1
Statement of the Hypothesis.....	3
METHOD	
Subjects.....	4
Instrument.....	4
Design and Procedure.....	4
RESULTS	
Scoring Procedures.....	7
Data Analysis Procedures.....	9
CONCLUSIONS.....	11
REFERENCES.....	12
APPENDIX A.....	13

LIST OF TABLES

	Page
Table 1 Means and Standard Deviations for the Written Exercises, Tests, and Choice Treatment Conditions on the Retention Test.....	9

INTRODUCTION

Statement and Background of the Problem

The purpose of this study was to determine whether required tests would result in greater retention of research concepts and procedures than required written exercises. The desirability of testing as a method of evaluation has been increasingly questioned. It has been suggested that the "cramming" phenomenon often associated with testing results in meaningless memorization of information which is quickly forgotten. It has also been stated that the anxiety generated by testing may actually have a detrimental, interfering effect on the intended learning. As a result of these beliefs, a trend toward utilization of alternative evaluation strategies has been developing.

This trend is especially visible in the competency-based education approach with its deemphasis of written tests (Houston, 1974; Johnson, 1972). While actual observation of the performance of competencies may indeed be the optimal evaluation strategy, not all concepts and procedures can be so evaluated (for example, Gallagher, 1973). There are a number of prerequisite concepts and procedures, acquisition of which facilitates achievement and demonstration of higher-order learnings (competencies), which do not lend themselves to performance evaluation. Even if they did, the large number of such prerequisites would

make individual performance evaluation infeasible. Since these concepts and procedures are prerequisites and therefore necessary, however, their acquisition must be assured in an effective and practicable manner.

If it is accepted that acquisition of such prerequisites must be assessed using written products as evidence, the problem then becomes one of determining which type of product, if any, results in greater retention of intended knowledge and skills. Reason would suggest that, due at least in part to the additional study time probably involved, the taking of tests should result in greater retention than the mere submission of written exercises. Although they may be philosophically distasteful, to students and educators alike, tests may be practically effective. The literature on the problem is meager. Researchers such as Andrews (1968), for example, have been more concerned with the validity and reliability data of various tests than with the effectiveness of different methods of assessment. A study by Calhoun (1972), however, did investigate the effectiveness of oral and written tests when utilized within a self-paced, individualized, instructional system. Calhoun's results indicated that oral and written testing were equally effective, and that both were more effective than no testing, when retention was measured with an instrument containing items parallel to those in the original tests.

Statement of the Hypothesis

There is research evidence to suggest that written testing represents an effective method of evaluation. Oral tests may be equally effective but their use is equally as impracticable as use of individual performance evaluation when dealing with large numbers of prerequisite concepts and procedures. Based on reason, Calhoun's study, and lack of evidence to the contrary, it was hypothesized that

students in an introductory educational research course who are required to reach criterion on each of a series of tests will exhibit significantly greater retention of the tested concepts and procedures on a final examination than students required to successfully complete a series of written exercises on the same concepts.

Both the tests and the final examination were composed of short-answer items and the final examination represented a representative sampling of items from each of the tests.

METHOD

Subjects

Subjects were 126 master's level students enrolled in the course Analysis and Application of Educational Research during the Fall quarter, 1976, at Florida International University (FIU), Miami, Florida. FIU is a commuter university located in a large, urban, multicultural area. Subjects were randomly assigned to one of three treatment conditions.

Instrument

The measure of the dependent variable, retention of research concepts and skills, was a 35-item short-answer test. The test was developed by selecting representative items from each of 10 previously administered criterion-referenced tests. The instrument concluded with several attitudinal questions related to participation in the study (see Appendix A).

Design and Procedure

At the beginning of the quarter, prior to the first class meeting, the 126 subjects were randomly assigned to one of three treatment conditions and to one of three instructors. Thus, each treatment group initially comprised 42 students and each instructor was assigned 42 students equally representing each of the 3 treatment conditions. No pretest was administered since pretesting in three previous quarters had shown that students entering the course are

completely lacking in research knowledge and skills.

Students in all treatment conditions received the same large-group lectures at the same time. The three instructors took turns being the featured lecturer of the week. Students in all treatment conditions were also required to demonstrate the same competencies, that is, perform the same tasks, which were evaluated by the students' respective instructors. Acquisition of subcompetencies, or prerequisite concepts and procedures, however, was demonstrated by different students in different ways, depending upon treatment condition:

Treatment 1 - Students demonstrated acquisition of all prerequisites through the submission of written exercises.

Treatment 2 - Students demonstrated acquisition of some prerequisites through submission of written exercises, but for most (about 80%), through the achievement of criterion on 10 short-answer tests.¹ The questions on the short-answer test were identical to the questions for which students in treatment 1 submitted written responses.

¹ Some prerequisites did not lead themselves to a testing situation. For example, the prerequisites for a task requiring the students to write a review of related literature involved the identification and annotation of at least 15 references related to a selected problem.

Treatment 3 - Students had their choice; for any given set of prerequisites they could submit them as written exercises or they could take a short-answer test.

Students in all groups had to achieve a criterion of 100% on both written exercises and tests. Students in all groups also recycled (resubmitted written exercises or retook tests) as many times as necessary in order to reach criterion. Complete self-pacing was permitted for all students; exercises were submitted and tests taken on dates and times individually determined by each student, within specified limits. All tests were administered, and all written exercises and tests were graded, by a graduate student who had previously taken the course and demonstrated a high level of proficiency. The graduate student was available 20 hours a week, including several afternoons and several evenings; students were free to come any time during posted hours in order to turn in exercises, take a test, or receive feedback on previous performance.

At the end of the quarter, during the last class meeting, all students were administered the "surprise" 35-item short-answer final examination as a posttest measure of retention. Students were asked to answer the questions to the best of their ability. They were told that although their scores would not affect their grades, their best effort would "contribute to research and to the improvement of the course."

RESULTS²

Scoring Procedures

As subjects were randomly assigned to treatment conditions using the list of students officially registered for the course, there were a number of students intended to be in the study who dropped the course before the first class meeting. This factor, coupled with normal attrition during the first few weeks of the quarter, resulted in a loss of subjects. Also, there were a number of students who, due to the self-pacing element, did not complete the course by the last class meeting. Only posttest scores of students who did complete the course by this date were included in data analyses. Resulting N's for the 3 treatment conditions were 16, 20, and 12 respectively.

All posttests were scored by the same graduate student who had previously scored all written exercises and tests. The scoring procedure was as follows:

1. A set of acceptable responses was developed for each item.
2. All test papers were alphabetized by last name, and first name if necessary.

² The authors wish to express their appreciation to Helen Diehl for her assistance in tabulating and analyzing the data.

3. Tests were scored 10 at a time, with correct and incorrect responses recored on a separate sheet. In order to avoid seeing the name of the student or identification of treatment condition, tests were scored backwards, last item first, first item last.
4. After a set of 10 tests was scored, 2 tests were randomly selected and rescored.
5. The remaining 8 tests from a set of 10 were rescored only on items 3, 10, 11, 14, 22, 23, and 30, those whose scoring was found to be most inconsistent.
6. The scorer took a break.
7. Steps 3-6 were repeated until all tests were completed.
8. One test was then randomly selected from each set of 10 (in one instance 8, since there were 48 tests), and completely rescored. Scorer error rate was estimated to be 1.6%.
9. All results were recorded on data sheets.
10. As a further assessment of scorer reliability, one of the instructors completely rescored all of the posttests. This was possible since the original scorer did not indicate correct or incorrect responses on the tests themselves. Inter-scorer reliability was calculated to be $\underline{r} = .94$.

Data Analysis Procedures

Means and standard deviations were calculated for each group (see Table 1).

Table 1

Means and Standard Deviations for
the Written Exercises, Tests, and
Choice Treatment Conditions
on the Retention Test

	Treatment Condition		
	Written Exercises	Tests	Choice
$\bar{X}$	16.50	20.70	14.33
SD	6.05	5.40	6.78
N	16	20	12

Note. Maximum score = 35.

A one-way analysis of variance indicated that there was a significant difference among the treatment conditions ($F = 4.44$, $df = 2/45$, $p < .05$). Application of Scheffé's multiple comparison procedure indicated that students in treatment condition 2, testing, retained significantly more concepts and skills than students in treatment condition 1, written exercises ($F = 4.10$, $df = 2/45$, $p < .05$), and students in treatment condition 3, choice, ($F = 7.79$, $df = 2/45$, $p < .05$). There was no significant difference

between the retention of students in treatment condition 1 and students in treatment condition 3 ($F = .25$, $df = 2/45$, $p < .05$). In terms of practical significance, students in treatment condition 2 actually retained an average of 25% more than students in treatment condition 1 ($\bar{X} = 20.70$ versus 16.50), and 44% more than students in treatment condition 3 ($\bar{X} = 20.70$ versus 14.33). It should be noted that not one student in treatment condition 3, choice, chose to take tests. Thus these students actually behaved essentially the same as students in treatment condition 1, written exercises.

Attitudinally, no student indicated an objection to participation in the study. More pertinent to the purpose of the study was the finding that many students in treatment condition 2 felt that although they had had to work harder, they had learned more.

CONCLUSIONS

The results of this study supported the original hypothesis. Students in an introductory educational research course who were required to reach criterion on each of a series of tests did exhibit significantly greater retention of the tested concepts and procedures on a final examination than students required to successfully complete written exercises on the same concepts.

The findings of this study suggest that formal written testing as a method of evaluation, no matter how distasteful it may be to students and educators, does promote greater retention of learned concepts and procedures. Consequently, the current trend toward the elimination of formal, written tests, especially in competency-based programs, may actually be counterproductive to one of education's major goals, student retention of learned concepts.

REFERENCES

- Andrews, A. S. Multiple choice and essay tests. Improving College and University Teaching, 1968, 16, 61-66.
- Calhoun, J. F. Oral vs. written testing in self-paced individualized instruction. Education, 1974, 94, 242-248.
- Houston, W. R. Competency based education. Exploring Competency Based Education, 1974, 5-15.
- Gallagher, P. D. Module format and the instructional process. Performance-Based Teacher Education, 1973, 2, 3-4.
- Johnson, C. E. Competency Based and Common Teacher Education Programs Compared. Athens: University of Georgia, College of Education, Competency Based Education Center, 1972, 7-10.

APPENDIX A
EDU 507 POSTTEST

Name _____

Your instructor was: Gay _____ Gallagher _____ Greenberg _____

You were in: Group 1 - All enablers written _____
Group 2 - Enabler Tests _____
Group 3 - Choice _____

Please answer the following questions to the best of your ability. Your score will not affect your grade but your best effort will contribute to research and to the improvement of the course.

Thank you for your cooperation.

1. Define or describe random sampling.
2. Define or describe "stratified sampling."
3. Define or describe "content validity."
4. Define or describe "concurrent validity."
5. List the procedures for determining predictive validity for the following situation:
You want to predict success in physics and you want to determine the predictive validity of a physics aptitude test.

6. Define or describe "reliability."

7. List or describe procedures for determining equivalent forms reliability.
8. Define or describe "primary sources" of data.
9. Define or describe "internal criticism" of data.
10. List three major ways to collect descriptive data.
11. List the major steps involved in conducting a questionnaire study.
12. List the major steps involved in conducting an observational study.
13. Discuss how a correlation coefficient is interpreted.
14. List one advantage and one disadvantage of correlational research.
15. Diagram and describe the basic causal-comparative design.
16. List one advantage and one disadvantage of causal-comparative research.
17. Define or describe "experimental research."

18. Define or describe "internal validity" and name one threat to internal validity.
19. Define or describe "external validity" and name one threat to external validity.
20. Define or describe "independent variable."
21. Define or describe "dependent variable."
22. Define or describe "extraneous variable."
23. Diagram and describe the pretest-posttest control group design.
24. Diagram and describe the non-equivalent control group design.
25. Which design is more valid, the pretest-posttest control group design or the non-equivalent control group design? Why?
26. Give two examples of interval data.
27. Define or describe the "mean" of a set of scores.

28. Define or describe the "median" of a set of scores.
29. Define or describe "Pearson r."
30. Define or describe "null hypothesis."
31. Define or describe "type I error."
32. Define or describe "level of significance."
33. Define or describe "independent samples."
34. Define or describe "t-test for non-independent samples," ie. tell when it is used.
35. Define or describe "chi square," ie. tell when it is used.

How did you feel about being involved in a study?

How did you feel about being in the group you were in (ie. written enablers versus enabler tests versus choice)?

Any other comments? Criticisms? Suggestions?